2014

A pseudo-likelihood analysis for incomplete warranty data with a time usage rate variable and production counts

Yu Qiu
Iowa State University, yuqiu@iastate.edu

Danial J. Nordman
Iowa State University, dnordman@iastate.edu

Stephen B. Vardeman
Iowa State University, vardeman@iastate.edu

This is an Accepted Manuscript of an article published by Taylor & Francis in IIE Transactions on February 8, 2013, available online: http://www.tandfonline.com/10.1080/0740817X.2013.770185

A Pseudo-Likelihood Analysis for Incomplete Warranty Data with a Time Usage Rate Variable and Production Counts

Yu Qiu, Daniel J. Nordman

Department of Statistics, Iowa State University

Stephen B. Vardeman

Department of Statistics, Department of Industrial and Manufacturing Systems Engineering, Iowa State University

Abstract: The most direct purpose of collecting warranty data is tracking associated costs. But they are also useful for quantifying a relationship between use rate and product time-to-first-failure and for estimating the distribution of product time-to-first-failure (which we model as depending upon use rate and a unit potential life length under continuous use). Employing warranty data for such reliability analysis purposes is typically complicated by the fact that some parts of some warranty data records are missing. We introduce pseudo-likelihood methodology for dealing with some kinds of incomplete warranty data (like those available in a motivating real case from a machine manufacturer). Based on this, we estimate a use rate distribution, the distribution of time to first failure, and the time associated with a cumulative probability of first failure.

Keywords: delivery time, failure times, reliability, repair time, use rate

1 Introduction

Manufacturers often compile data on products which fail during a warranty period. The resulting warranty database can, in principle, be applied to study the reliability of products, improve product quality, and adjust future policies for warranty coverage. However, the statistical analysis of warranty data presents some difficult challenges. The main complication is that typically information on failure times and other product characteristics is available only for units in the warranty database (i.e., units failing during a warranty

period). In important papers on the analysis of warranty data, Lawless (1998) and Karim and Suzuki(2005) provided overviews of several approaches for estimating warranty costs and failure time distributions. As suggested in those works, inference based on warranty databases has received increasing attention over the past 25 years (cf. Suzuki (1985ab), Kalbfleisch, Lawless and Robinson (1991), Lawless, Hu and Cao (1995), Kalbfleisch and Lawless (1988a,1996), Hu and Lawless (1996ab, 1997), Wu and Meeker (2002), Alam and Suzuki (2009)); see Lawless (1998) and Karim and Suzuki (2005) for further references. However, we have found that existing analysis methods typically require records in a warranty database to be complete. This is not always possible or realistic. That is, in addition of lacking information for (non-failing) units naturally outside warranty database, the records in a warranty database can also be incomplete and exhibit a range of messy types of missing information. We present some possibilities in this regard with respect to a real warranty database problem and propose methodology for dealing with these issues.

The motivation for our work comes from a machine manufacturer and concerns an electronic assembly. For units requiring service within a one-year guarantee period, the information available in the real warranty database ideally includes: serial numbers, assembly (i.e., manufacture) time, delivery time (i.e., field introduction upon sale), first repair time and total running time (in hours) before first repair. Among these "times," running time is a time length, while the others denote points in time. To be consistent in the following analysis, we express all times in units of months. We are interested in a unit's running time before repair and the calendar time between a unit's delivery and repair. Ideally, the first is directly available in the warranty database while the second is the difference between delivery and repair times (both of which are also ideally found in the warranty database). But not all the records in the real warranty database are complete. For varied unknown reasons (that are hopefully unrelated to unit failure histories), some of the units represented in the warranty database have missing information. For example, some records lack delivery or repair times, while other records fail to include running times. We summarize all the possible cases of complete and incomplete warranty records in Table 1.

In addition to the warranty database, also available in our motivating case are production

records giving the number of units assembled each month prior to the close of data collection (30 months in total).

The setting of appropriate warranty periods and use of warranty data of the type described here create a variety of interesting and important statistical problems. This paper focuses on two of these. The first is to estimate the distributions of failure time (as the actual time difference between delivery and repair, or as a theoretical time until failure under continuous running) and overall usage rate (the fraction of the time until failure that a unit is actually used). In Section 2, we propose a class of probability models that link calendar failure times, failure times under continuous running, and usage rates. In Section 3

Table 1: All warranty record types. (An "X" indicates a value available in the warranty database. Running time is a time length, the remaining time variables are points in time.)

Assembly Time	Delivery Time	Repair Time	Running Time	Case Type
X	X	X	X	1
X	X	X		2
X	X		X	3
X	X			4
X		X	X	5
X		X		6
X			X	7
X				8
	X	X	X	9
	X	X		10
	X		X	11
	X			12
		X	X	13
		X		14
			X	15
				16

we formulate pseudo-likelihood methods to find estimates of model parameters in a manner which accounts for varieties of missing information. These could, for example, be the basis for comparisons across different product groups.

After considering estimation of parameters, the next concern is the estimation of probability of failure by a given time of service and the estimation of unit life length at a given cumulative failure probability (that is important for judging how to set warranty policy). We identify a methodology for extending inference beyond parameters to these important functions of model parameters. In addition to providing point estimators of parameters in failure time and usage rate distributions, the pseudo-likelihood approach also provides standard errors to quantify the precision of the point estimators and functions of them. Based on these, confidence limits can be provided for both parameters and parametric functions.

In Section 4, we discuss the simulation of databases (with known parameters) more or less consistent with our motivating case. We then use such simulated databases to examine the effectiveness of our methodology. In Section 5, we apply our complete methodology to a single simulated data set and illustrate the practical inferences that are possible. We conclude by mentioning some possible extensions of the present work in Section 6.

Having outlined the structure of our warranty data and related inference issues, we end this section by describing how our methodology fits into some existing literature for handling warranty databases. In particular, there exist connections between our inference setting and scenarios considered by Lawless, Hu and Cao (1995), Hu and Lawless (1996ab, 1997) and Lawless (1998). Those works similarly considered fitting parametric models for failure time distributions based on warranty data. But they also assumed that complete information was available on failure times and other covariates for units in the warranty database. While covariate information (e.g., assembly or delivery times) is also available in our motivating warranty data, our warranty records are themselves incomplete and can lack both failure times (i.e., "repair minus delivery" times) and other covariate values to varying degrees as named in Table 1. Also, in the works mentioned above, the parametric models involved were intended only to describe failure times (possibly specified conditionally on covariates), while Section 2 develops probability models which directly and simultaneously describe failure

time, overall usage rate, and actual running time. The latter is a variable directly available in our database (though potentially missing for some units) with no immediate counterpart in the previous work (e.g., analysis of auto warranty claims from Hu and Lawless (1996ab, 1997)). Finally, the existing works above developed a pseudo-likelihood for parametric inference, where the pseudo-likelihood arose when attempting to incorporate non-failing units (i.e., units not in the warranty database) into likelihood inference based on a partial sample of non-failed units (i.e, to collect covariate values on these in addition to the warranty database). Our pseudo-likelihood arises out of a similar need to quantify the informational contribution of non-failed units, but we have no detailed information on units outside of the warranty database, only totals of assembled units over each month of data collection. We consequently provide a different formulation of a pseudo-likelihood based on the warranty database and the monthly production totals, as described in Section 3. Additionally, it is important to note that likelihood approaches of Philips and Sweeting (2001) and Alam and Suzuki (2009), which require distributional assumptions on how non-warranty database units are "censored" (assumptions difficult to verify based on warranty data alone), are not directly applicable here. These methods use counts of non-failed units, all of which are assumed to be in service (i.e., sold). In contrast, our production counts available in addition to the warranty database only roughly suggest when non-failed units are assembled, not sold into service, and this information is complicated by the fact that we lack assembly dates for some units in the warranty database. Consequently as part of our pseudo-likelihood method in Section 3, we estimate how much of each month's production has not failed and when these units are delivered into service, based on patterns in the warranty data.

2 Modeling

We need to develop some notation to describe the variables of interest and probability models for these. For a given unit, define a random variable C, the "unit's failure time," as the difference between delivery and first repair time . Also let random variable A be the "unit's actual use time," the unit's running time until first failure. It holds that $C \geq A$ and a positive difference $C - A$ allows the possibility that there are periods of non-use among

periods of operation of a unit. As determined by the company's warranty plan, the warranty period is one year (12 months) in terms of C.

We aim first to develop a model for the joint distribution of (A, C). To this end, we introduce two other variables for which model formulation is straightforward. For a given unit, let

$$U = A/C \in [0, 1], \tag{1}$$

denote the unit's usage rate and let the continuous variable T denote the unit's running time until failure under a 100% use condition. The time T represents a theoretical and unobservable variable related to a unit's potential. A reasonable model for the relationship between the theoretical failure time T and the actual running time A is

$$T = U^{\theta} A = C^{-\theta} A^{1+\theta}, \tag{2}$$

where θ is a real-valued parameter. This formulation provides some important possible interpretations in terms of θ. For $\theta > 0$, a fractional usage rate U (less than 1) increases the actual running time A above the running time T under 100% use, while A will be less than T for $\theta < 0$. Hence, the sign and magnitude of θ can suggest what failure modes are operating (e.g. wear-out from continuous running or failure produced by frequent on/off switches).

A model for the joint distribution of (A, C) for a given unit can be formulated from assumptions which can be most readily framed in terms of variables T and U. For any unit, suppose that T has a lognormal(μ, σ^2) distribution with density

$$f_T(t|\mu, \sigma^2) = \frac{1}{t\sigma\sqrt{2\pi}} \exp(-\frac{(\ln t - \mu)^2}{2\sigma^2}), \qquad t > 0$$

involving location and scale parameters μ and σ^2 for $\log T$. This is a common and flexible failure-time model for describing lifetimes under continuous use (see Chapter 4 of Meeker and Escobar, 1998). Also, suppose that the usage rate U has a beta(α, β) distribution with density

$$f_U(u|\alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} u^{\alpha-1}(1-u)^{\beta-1} \quad \text{for} \quad 0 < u < 1$$

with shape parameters $\alpha > 0, \beta > 0$, where

$$\mathrm{B}(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\mathrm{d}u = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

in terms of the gamma function $\Gamma(\cdot)$. The beta distribution is a flexible model for describing proportions such as U. Finally, for any given unit, assume the potential running time T under 100% use is independent of the usage rate U.

Under these assumptions, the joint density of (T, U) for a given unit can be easily translated into a joint density for the observable values (a, c) of variables (A, C) given by

$$f_{A,C}(a, c|\mu, \sigma^2, \alpha, \beta, \theta) = f_T(c^{-\theta}a^{1+\theta}) f_U(c^{-1}a) c^{-2-\theta} a^{1+\theta}, \quad 0 < a < c. \tag{3}$$

From this, the marginal densities of times A and C can be derived as

$$\begin{aligned} f_C(c|\mu, \sigma^2, \alpha, \beta, \theta) &= \int_0^c f_{A,C}(a, c|\mu, \sigma^2, \alpha, \beta, \theta) da \\ &= \int_0^1 f_T(cu^{1+\theta}|\mu, \sigma^2) f_U(u|\alpha, \beta) u^{1+\theta} du \end{aligned} \tag{4}$$

and

$$\begin{aligned} f_A(a|\mu, \sigma^2, \alpha, \beta, \theta) &= \int_a^\infty f_{A,C}(a, c|\mu, \sigma^2, \alpha, \beta, \theta) dc \\ &= \int_0^1 f_T(u^\theta a|\mu, \sigma^2) f_U(u|\alpha, \beta) u^\theta du. \end{aligned} \tag{5}$$

We note for future reference that while there are no closed forms for these marginal densities, particular values of these functions can be obtained from expressions (4) and (5) via numerical integration.

In this framework, it is also possible to derive marginal distributions for other potentially useful variables and joint distributions for pairs of variables, such as the length of ownership and usage rate pair (C, U). However, for the purpose of model fitting, it is most useful and relevant to consider the distribution of (A, C) for each unit, corresponding to values potentially obtainable from the database. An important point is that, as mentioned in the Introduction, this framework gives parametric probability models for lengths of ownership C and running times A to failure in addition to usage rates U, which are not immediate from models in previous warranty analyses (cf. Lawless, 1998). An advantage of the present approach is that parameters values $(\mu, \sigma^2, \alpha, \beta, \theta)$ allow easy description and interpretation of life length models and ones conditioned on usage rate, as explained next.

Note additionally that

$$C = \frac{T}{U^{\theta+1}}$$

and therefore, conditional on usage rate U, the calendar time C that a unit operates before failure has a lognormal distribution,

$$C|U \sim \text{lognormal}(\mu - (\theta+1)\ln U, \sigma^2).$$

From this, the probability $F_C(t|\mu,\sigma^2,\alpha,\beta,\theta)$ of unit failure by a given time $t > 0$ after delivery may be expressed as

$$\begin{aligned} F_C(t|\mu,\sigma^2,\alpha,\beta,\theta) &\equiv P(C \le t|\mu,\sigma^2,\alpha,\beta,\theta) \\ &= \int_0^1 \Phi\left(\frac{\ln(t)-\mu+(\theta+1)\ln u}{\sigma}\right) f_U(u|\alpha,\beta)du \end{aligned} \tag{6}$$

in terms of the usage rate density f_U and the standard normal cumulative distribution function $\Phi(x) = \int_{-\infty}^{x} e^{-y^2/2}/\sqrt{2\pi}dy$, $x \in \mathbb{R}$. (Again, while there is no closed form for this cumulative probability, for particular t the integral in display (6) can be computed numerically.) Inference about (6) is important in our motivating example and will be considered later.

Additionally, the conditional log-normal distribution of C implies that, given a unit's usage rate U, the mean value of C on the log scale shifts by an increment depending on log usage rate and the parameter $\theta \in \mathbb{R}$. Similarly, given U, the actual running time A is also conditionally lognormal$(\mu+\theta\ln U, \sigma^2)$ with the same shape parameter σ^2 as that of the distribution for T. Because $0 < U < 1$, the conditional distributions of A and C given U have means (on the log-scale) with behavior depending on θ. The conditional distributions of both $C|U$ and $A|U$ have log-scale mean values smaller than that of the distribution of T when $-1 < \theta < 0$; when $\theta < -1$, the conditional distribution of $C|U$ has an log-scale mean larger than that of the distribution of T, while the conditional distribution of $A|U$ has a log-scale mean smaller than that of the distribution of T; and the conditional distributions of $C|U$ has a smaller log-scale mean than that of the distribution of T, while the conditional distribution of $A|U$ has a larger log-scale means than that of the distribution of T when $\theta > 0$.

Next we can develop a log-pseudo-likelihood (in the future we will abbreviate pseudo-likelihood as PL) function corresponding to the available data for the purpose of estimating the unknown parameters $(\mu,\sigma^2,\alpha,\beta,\theta)$ by "maximum PL." The term "pseudo" is used here

because a usual likelihood function is not directly possible in the presence of the incomplete nature of the motivating warranty database. The PL attempts to be an approximation of the likelihood which would result from complete information.

3 PL Inference Methodology

We develop inference through an appropriate log-PL function that is a sum of log-PL terms, one for each unit produced, including those not represented in the warranty database but appearing in a separate data set of assembly counts. Such log-PL terms will express the informational contribution of each unit for estimating the model parameters $\mu, \sigma, \alpha, \beta, \theta$. The form appropriate for each contribution to the PL depends upon whether the unit has appeared in the warranty database and what information about the unit is available there. Recall that Table 1 provides a listing and naming convention for the kinds of cases that could potentially appear in the warranty database.

In Section 3.1, we first develop the log-PL contributions for those units in the warranty database (failing during the 12-month warranty period and possibly having missing information). For these warranty database units, their log-PL contributions would perhaps be better described as "true log-likelihood" contributions, meaning that we do in fact specify an exact probability of observing each unit in terms of the probability models of Section 2 and their parameters $\mu, \sigma, \alpha, \beta, \theta$ (i.e., the traditional sense of likelihood). However, we also need to include the log-PL contributions for units which have not failed under warranty, which is considered in Section 3.2 based on the monthly assembly counts in addition to the warranty database. Because non-failing units lack failure information, we can only approximate their probability contributions (i.e., these cannot be specified purely in terms of model probabilities but require some additional estimation steps), which are then "pseudo" and not the usual log-likelihood contributions based purely on the models at hand. We shall clarify this point in Section 3.2. For simplicity, we refer to all units having a "log-PL contribution" and Section 3.3 describes some inference possibilities based on the final log-PL function.

3.1 Log-PL terms for units in the warranty database

In the following, we continue to denote the actual running time and calendar time to first failure of a given unit as (A, C) and assume that all units are independent. Additionally, we define some additional random variables for the use in the following. Let the "closing time" be the time when the manufacturer stops to collect data and "starting time" be the time when the manufacturer starts to collect data, and define the following variables for a given unit:

$$H \equiv \min(12, \text{difference between closing time and delivery time}), \tag{7}$$

$$S \equiv \min(12, \text{difference between repair time and assembly time}), \tag{8}$$

$$R \equiv \min(12, \text{difference between closing time and assembly time}), \tag{9}$$

and

$$Q \equiv \min(12, \text{difference between repair time and starting time}). \tag{10}$$

In considering a unit from the warranty database, the above variables become useful in formulating contributions to the log-PL, particularly when units lack information about actual running time and calendar time to first failure (A, C). Recall, however, that warranty database units vary in their level of missing information so that some variables above may not be observed for certain units. We suppose that values of (H, S, R, Q) for a unit, depending on the unit's delivery and assembly times as well as starting and closing times of data collection, are independent of the unit's calendar and running times to failure (A, C), which is a reasonable modeling assumption. The observed values of A, C, R, H, S, Q will be denoted as a, c, r, h, s, q in the following.

For each of the 16 informational cases in Table 1, we provide a unit's contribution to a log-PL function. There are in total 12 distinct types of contributions to the log-PL to be formulated (as some cases in Table 1 can be grouped and handled similarly). For clarity, we separately enumerate and describe the 12 "types" below, using subscripts ji to denote the ith unit in the jth type class, $j = 1, ..., 12$. We also let N_j denote the number of units available in the jth type class. (For example, the subscript $1i$ denotes unit i for the first type of contribution considered.)

1. For a case 1 or case 9 unit in Table 1 where delivery time, repair time and actual running time are known, we observe values (a_{1i}, c_{1i}) for each unit and suppose that there are N_1 such units involved. From the joint density (3) of (A, C), the contribution of these units to the log-PL function is then

$$\mathscr{L}_1(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=1}^{N_1} \ln f_{A,C}(a_{1i}, c_{1i}|\mu, \sigma^2, \alpha, \beta, \theta). \tag{11}$$

2. For a case 2 or case 10 unit in Table 1, the delivery time and repair time are known, so that a value c of calendar time C is observed, but actual running time A is not observed. Hence, we cannot use the form (11) for these units. However, we can use the marginal density (4) for the calendar time C. If there are N_2 such case 2 and 10 units with observed values c_{2i}, then their contribution to the log-PL is

$$\mathscr{L}_2(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=1}^{N_2} \ln f_C(c_{2i}|\mu, \sigma^2, \alpha, \beta, \theta).$$

3. For a case 15 unit in Table 1, a value a of the actual running time A is known, but calendar time C is not. We now use the marginal density (5) for the running time A (instead of C treated directly above). If there are N_3 such case 15 units with observed values a_{3i}, then their contribution to the log-PL is

$$\mathscr{L}_3(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=1}^{N_3} \ln f_A(a_{3i}|\mu, \sigma^2, \alpha, \beta, \theta).$$

4. For a case 3 or case 11 unit in Table 1, where the repair time is missing, we know the calendar time C should be at least the observed value a of running time A. Additionally, we know that the unit's calendar time C cannot be more than the observed value h of the variable H from (7), representing the minimum of 12 months (since only units failing within the warranty period are in the database) and the difference between the delivery time and the closing time. The probability contribution of such a unit to the PL is then

$$\begin{aligned} & f_A(a|\mu, \sigma^2, \alpha, \beta, \theta) P\Big(a < C < h|\mu, \sigma^2, \alpha, \beta, \theta\Big) \\ = & \int_a^h f_{A,C}(a, c|\mu, \sigma^2, \alpha, \beta, \theta) dc. \end{aligned}$$

If there are N_4 case 3 and 11 units with values a_{4i} and h_{4i}, then their contribution to the log-PL is

$$\mathscr{L}_4(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=1}^{N_4} \ln\Big(\int_{a_{4i}}^{h_{4i}} f_{A,C}(a_{4i}, c|\mu, \sigma^2, \alpha, \beta, \theta) dc \Big). \tag{12}$$

(The integrals here must be computed numerically.)

5. For a case 5 unit in Table 1, where the delivery month is missing, we again know the calendar time C should be at least the observed running time a and must be less than observed (available) value s of the variable S from (8), the minimum of 12 months and the difference between the repair and assembly times for the unit. If there are N_5 case 5 units available, with observed values a_{5i} and s_{5i}, then their contribution to the log-PL is

$$\mathscr{L}_5(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=1}^{N_5} \ln\Big(\int_{a_{5i}}^{s_{5i}} f_{A,C}(a_{5i}, c|\mu, \sigma^2, \alpha, \beta, \theta) dc \Big),$$

analogously to (12).

6. For a case 7 unit in Table 1, where both the delivery time and repair time are missing, we know the calendar time C should be at least the observed running time a and must be less than observed value r of the variable R from (9), the minimum of 12 months and the difference between the assembly and closing times. If there are N_6 case 7 units available with observed values a_{6i} and r_{6i}, then their contribution to the log-PL is

$$\mathscr{L}_6(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=1}^{N_6} \ln\Big(\int_{a_{6i}}^{r_{6i}} f_{A,C}(a_{6i}, c|\mu, \sigma^2, \alpha, \beta, \theta) dc \Big). \tag{13}$$

7. For a case 13 unit in Table 1, where both delivery time and assembly time are missing, we know the calendar time C should be at least the observed running time a and less than the observed value q of Q from (10), representing the minimum of 12 months and the difference between the repair and starting times. If there are N_7 case 13 units available with observed values a_{7i} and q_{7i}, then their contribution to the log-PL is

$$\mathscr{L}_7(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=1}^{N_7} \ln\Big(\int_{a_{7i}}^{q_{7i}} f_{A,C}(a_{7i}, c|\mu, \sigma^2, \alpha, \beta, \theta) dc \Big). \tag{14}$$

8. For a case 4 or 12 unit in Table 1, both repair time and running time are missing. For such a unit, we do not know the value of A and only know that the calendar time C is less than the observed h value of the variable H from (7), representing the minimum of 12 months and the difference between the delivery and closing times. If there are N_8 case 4 and 12 units available with observed values h_{8i}, then their contribution to the log-PL is

$$\mathscr{L}_8(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=1}^{N_8} \ln F_C(h_{8i}|\mu, \sigma^2, \alpha, \beta, \theta), \tag{15}$$

using the cumulative failure distribution F_C of C from (6).

9. For a case 6 unit in Table 1, where both delivery time and running time are missing, we do not know the value of A and only know that the calendar time C is less than the observed value s of S from (8), representing the minimum of 12 months and the difference between the repair and assembly times. If there are N_9 case 6 units available with observed values s_{9i}, then their contribution to the log-PL is

$$\mathscr{L}_9(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=1}^{N_9} \ln F_C(s_{9i}|\mu, \sigma^2, \alpha, \beta, \theta)$$

computed analogously to (15).

10. For a case 14 unit in Table 1, where only repair time is known, we only know that the calendar time C is less than the observed value q of the variable Q from (10), representing the minimum of 12 months and the difference between the repair and starting times. If there are N_{10} case 14 units available with observed values of q_{10i} from (10), then their contribution to the log-PL is

$$\mathscr{L}_{10}(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=1}^{N_{10}} \ln F_C(q_{10i}|\mu, \sigma^2, \alpha, \beta, \theta).$$

This is the version of (14) where values for A are unknown.

11. For a case 8 unit in Table 1, where only the assembly time is known, we only know that the calendar time C is less than the observed value r of the variable R from (9), representing the minimum of 12 months and the difference between the closing and

assembly times. If there are N_{11} such units available with observed values of r_{11i}, then their contribution to the log-PL is

$$\mathscr{L}_{11}(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=1}^{N_{11}} \ln F_C(r_{11i}|\mu, \sigma^2, \alpha, \beta, \theta).$$

This is the version of (13) where values for A are unknown.

12. For a case 16 unit in Table 1, none of the four variables ideally in the database is known. The only information available is that (assuming that data collection spans at least 12 months) the unit failed within the warranty 12 month period. If there are N_{12} case 16 units, then their contribution to the log-PL is

$$\mathscr{L}_{12}(\mu, \sigma^2, \alpha, \beta, \theta) = N_{12} \ln F_C(12|\mu, \sigma^2, \alpha, \beta, \theta).$$

3.2 Log-PL terms for units not in the warranty database

As mentioned in Section 1, there are 30 months of production in the motivating example. Let M_i denote the number of units assembled in month i and let M_i^* denote the number of units assembled in month i which are *not* in the warranty database. Here we index months as $i = 0, 1, 2, ..., 30$ and month 0 is the closing month (the month when the manufacturer stops collecting data), month 1 is the month immediately before the closing month, and so on. For each month of assembly i, we need to incorporate the M_i^* non-failed units into inference about failure time models and parameters. To not do so would cause bias in estimation and misleading inference, especially when large numbers of assembled units M_i^* do not fall into the warranty database (cf. Kalbfleisch & Lawless, 1988a,1989).

However, the first complication is that we lack information for placing some units in the warranty database (which have missing assembly information) into their corresponding assembly months. That is, while total counts M_i are available each month, we do not know how many units in the warranty database were assembled in month i, namely $M_i - M_i^*$, and therefore do not know the number of non-failed units M_i^* for each assembly month. Hence, we need to form *estimators* $\hat{M}_i^*$ of the unknown counts, as described in Section 3.2.1. The second complication is that even if the monthly assembly counts M_i^* for units not in the warranty database were known, delivery times are unknown for these units. This means we

do not know exactly when the units are sold (if at all) or placed into service. Forms for PL terms for units not in the warranty database must accommodate this lack of information. Consequently, we are forced to *estimate* the probability contribution, say $p_i(\mu, \sigma^2, \alpha, \beta, \theta)$, of a non-failed unit assembled in month i, which depends on the parameters. An estimator $\hat{p}_i(\mu, \sigma^2, \alpha, \beta, \theta)$ is described in Section 3.2.2. We then let $\ln \hat{p}_i(\mu, \sigma^2, \alpha, \beta, \theta)$ denote the log-PL contribution for each of the $\hat{M}_i^*$ non-failed units estimated to be assembled in month i, $i = 0, \ldots, 30$, and the overall contribution to the log-PL becomes

$$\mathcal{L}^*(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=0}^{30} \hat{M}_i^* \ln \hat{p}_i(\mu, \sigma^2, \alpha, \beta, \theta) \tag{16}$$

for all units *not* in the warranty database. Finally, $\mathcal{L}^*(\mu, \sigma^2, \alpha, \beta, \theta)$ is added to warranty data-based log-PL contributions from Section 3.1 to produce a final log-PL function for inference in Section 3.3. To re-iterate our discussion at the beginning of Section 3, we note the distinction that estimation of a probability term (e.g., $\ln \hat{p}_i(\mu, \sigma^2, \alpha, \beta, \theta)$ for a non-warranty database unit) creates true pseudo-likelihood contribution, which differs from a "true likelihood" contribution (for warranty database units) described in Section 3.1 (where probabilities are stated in terms parameters but are not estimated). However, we continue to refer to all units as having a "log-PL contribution" for simplicity.

3.2.1 Estimation of monthly assembly counts for non-failed units

For case 1-8 units, assembly months are known, so we can directly subtract their counts from the assembly counts M_i of the corresponding months. But we cannot make similar adjustments to account for the other case 9-16 units represented in the warranty database (due to missing assembly information). Rather, we can only estimate appropriate adjustments, based on patterns observed in the warranty database under the assumption that the pattern can be extended to units not in the warranty database.

For each month $i = 0, 1, 2, ..., 30$, we compute an estimator $\hat{M}_i^*$ of the number of units assembled in month i and not in the warranty database M_i^* as

$$\hat{M}_i^* = M_i - M_{i,\text{1-8}} - \hat{M}_{i,\text{9-12}} - \hat{M}_{i,\text{13-14}} - \hat{M}_{i,\text{15-16}}$$

where

$$M_{i,\text{1-8}} = \text{ number of case 1-8 units assembled in month } i,$$

$$\hat{M}_{i,\text{9-12}} = \text{ estimated number of case 9-12 units assembled in month } i,$$

$$\hat{M}_{i,\text{13-14}} = \text{ estimated number of case 13-14 units assembled in month } i,$$

and

$$\hat{M}_{i,\text{15-16}} = \text{ estimated number of case 15-16 units assembled in month } i.$$

Since case 9-16 units in the warranty database lack assembly times, we formulate the estimators $\hat{M}_{i,\text{9-12}}, \hat{M}_{i,\text{13-14}}, \hat{M}_{i,\text{15-16}}$ by a process of careful "matching" against other warranty cases which do have assembly month information, as developed next.

For case 9-12 units, we do not know the assembly months. Instead, only the delivery months are known. So to estimate the assembly month for each 9-12 case unit, we first look to case 1-4 units in Table 1 which have both the assembly months and delivery months, and get the counts $n_{ij} \equiv$ number of case 1-4 units delivered in month j and assembled in month i, $0 \le j \le i \le 30$ and consider the fraction

$$b_{ij} = \frac{n_{ij}}{n_{jj} + n_{(j+1)j} + n_{(j+2)j} + \cdots + n_{30j}} \tag{17}$$

as an estimate of the proportion of case 9-12 units delivered in month j which are assembled in month i. Then an estimated number of case 9-12 units assembled in month i is

$$\hat{M}_{i,\text{9-12}} = \sum_{j=0}^{i} \left[b_{ij} \times \begin{array}{l} \text{number of case 9-12 units in warranty} \\ \text{database delivered in month } j \end{array} \right].$$

For case 13 and 14 units, we similarly look to case 1-2 and case 5-6 units which have both the assembly month and repair month and find counts $m_{ij} \equiv$ number of case 1-2 and 5-6 units repaired in month j and assembled in month i, $0 \le j \le i \le 30$, and consider the fraction

$$c_{ij} = \frac{m_{ij}}{m_{jj} + m_{(j+1)j} + m_{(j+2)j} + \cdots + m_{30j}}$$

as an estimate of the proportion of case 13-14 units repaired in month j which are assembled in month i. Then an estimated number of case 13-14 units assembled in month i is

$$\hat{M}_{i,\text{13-14}} = \sum_{j=0}^{i} \left[c_{ij} \times \begin{array}{l} \text{number of case 13-14 units in warranty} \\ \text{database repaired in month } j \end{array} \right].$$

For case 15 and 16 units, we look to the case 1-8 units and determine counts $l_i \equiv$ number of units assembled in month i, $0 \leq i \leq 30$ and consider the fraction

$$d_i = \frac{l_i}{l_0 + l_1 + l_2 + \cdots + l_{30}}$$

as an estimate of the proportion of case 15-16 units assembled in month i in the warranty database. Then estimated number of case 15-16 units assembled in month i is

$$\hat{M}_{i,\text{15-16}} = d_i \times \text{number of case 15-16 units in warranty database.}$$

What we have suggested above is using estimates for the expected values of counts in place of unavailable observed counts. A more sophisticated analysis might assign probabilities to each possible configuration of how units without assembly months are distributed and use those in a likelihood term. But the simpler analysis suggested here is adequate for our present purposes, and we shall not pursue this more complicated possibility.

3.2.2 Formulation of a PL-contribution for non-failed units

The above process leaves us with adjusted counts $\hat{M}_i^*$ of units assembled in month i where these units do not fail under the 12 month warranty period and still function at either the end of their warranty periods or the closing time for data collection. Then we need to find an appropriate term for these non-failed units for entry into the log-PL function. However, a further complication is that we are missing the delivery months for these non-failed units, so we do not exactly know when they are sold, if at all.

Consider those units assembled in a particular month but not in the warranty database (i.e., not yet accounted for in the log-PL function). Suppose that we had the probabilities

$$\nu_{ij} = \text{probability that a unit assembled in month } i \text{ is delivered in month } j$$

for $0 \leq j \leq i \leq 30$ and that $\sum_{j=0}^{i} \nu_{ij} = 1$; the latter condition implies that a unit assembled in month i will be sold/delivered among the months $0 \leq j \leq i$ and we will discuss this

condition in more detail at the end of this section. In this case, we could represent the probability, say $p_i(\mu, \sigma^2, \beta, \alpha)$, that a given unit assembled in month i would not appear in the warranty database as

$$p_i(\mu, \sigma^2, \beta, \alpha) \equiv \sum_{j=0}^{i} \nu_{ij} P(C > \min(j, 12) | \mu, \sigma^2, \beta, \alpha), \tag{18}$$

recalling that 12 months is the warranty period. Again, we might then use $\hat{M}_i^*$ times $\ln p_i(\mu, \sigma^2, \beta, \alpha)$ to represent the contribution of month i's assemblies not appearing in the warranty database to the log-PL function. However, we cannot obtain or formulate the probabilities (18) directly, because the v_{ij} terms are unknown and cannot be expressed in terms of the probability models for failure time or usage rate developed in Section 2. The only information available about the distribution of times between assembly and delivery is in the warranty database, which we use to estimate v_{ij} terms and form an estimate or approximation $\hat{p}_i(\mu, \sigma^2, \beta, \alpha)$ of (18).

Analogously to (17), if we define

$$e_{ij} = \frac{n_{ij}}{n_{i0} + n_{i1} + \cdots + n_{ii}}$$

as an estimate of the proportion of warranty database units assembled in month i which are delivered in month j, then roughly

$$e_{ij} \approx \frac{\nu_{ij} F_C(t_j^* | \mu, \sigma^2, \alpha, \beta, \theta)}{\sum_{j=0}^{i} \nu_{ij} F_C(t_j^* | \mu, \sigma^2, \alpha, \beta, \theta)}, \quad \text{for } t_j^* = \min(j, 12),$$

as the right-hand side represents the conditional probability of a unit being delivered in the month j given that it fails under warranty and is assembled in month i. Recall that $F_C(t|\mu, \sigma^2, \alpha, \beta, \theta) = P(C \leq t|\mu, \sigma^2, \alpha, \beta, \theta) = 1 - P(C > t|\mu, \sigma^2, \alpha, \beta, \theta)$ for $t > 0$. Since the denominator is a fixed number, we can expect that approximately $e_{ij} \propto \nu_{ij} F_C(t_j^* | \mu, \sigma^2, \alpha, \beta, \theta)$ or reciprocally,

$$\nu_{ij} \propto \frac{e_{ij}}{F_C(t_j^* | \mu, \sigma^2, \alpha, \beta, \theta)}$$

so that upon normalization

$$\nu_{ij} \approx \frac{\frac{e_{ij}}{F_C(t_j^* | \mu, \sigma^2, \alpha, \beta, \theta)}}{\sum_{j=0}^{i} \frac{e_{ij}}{F_C(t_j^* | \mu, \sigma^2, \alpha, \beta, \theta)}}.$$

Using this data-based approximation for v_{ij} terms, we obtain an estimate $\hat{p}_i(\mu, \sigma^2, \alpha, \beta, \theta)$ of (18) given by

$$\begin{aligned}\hat{p}_i(\mu, \sigma^2, \alpha, \beta, \theta) &= \sum_{j=0}^{i} \frac{e_{ij} \frac{1-F_C(t_j^*|\mu,\sigma^2,\alpha,\beta,\theta)}{F_C(t_j^*|\mu,\sigma^2,\alpha,\beta,\theta)}}{\sum_{j=0}^{i} \frac{e_{ij}}{F_C(t_j^*|\mu,\sigma^2,\alpha,\beta,\theta)}} \\ &= 1 - \frac{1}{\sum_{j=0}^{i} \frac{e_{ij}}{F_C(t_j^*|\mu,\sigma^2,\alpha,\beta,\theta)}},\end{aligned}$$

(using the fact that $\sum_{j=0}^{i} e_{ij} = 1$). A sensible term to represent the contribution of a non-failed unit assembled in month i in the log-PL is then $\ln \hat{p}_i(\mu, \sigma^2, \alpha, \beta, \theta)$. Recall from Section 3.2.1 that $\hat{M}_i^*$ is an estimate of the number of non-failed unit assembled in month i. By weighting these estimated counts by their approximate log-PL contributions, we obtain a final version of (16) as

$$\mathcal{L}^*(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=0}^{30} \hat{M}_i^* \ln \left(1 - \frac{1}{\sum_{j=0}^{i} \frac{b_{ij}}{F_C(t_j^*|\mu,\sigma^2,\alpha,\beta,\theta)}} \right), \tag{19}$$

to represent the log-PL contribution of all units not in the warranty database. As previously, F_C is computed numerically from (6).

We note that the log-PL component in (19) roughly resembles a formulation proposed by Lawless (1998), whereby the probability contribution of non-failed units is specified conditionally in terms of covariates whose distribution must then be estimated or empirically determined. In our framework, a similar probability (18) is stated conditionally in terms of a non-failed unit's assembly and delivery months with a corresponding distribution (i.e., v_{ij} terms above) requiring estimation. Unlike the Lawless framework however, we do not have a supplementary sample of non-failed units to inform estimation of the assembly/delivery month distribution, and so we use information solely within the warranty database. Recall that the condition $\sum_{j=0}^{i} \nu_{ij} = 1$ used above to formulate (18) implies that a unit assembled in month $i = 0, \ldots, 30$ will be sold/delivered among the months $0 \leq j \leq i$. This condition is not strictly met for units outside of the warranty database, which may be delivered after the closing time of data collection. But, this is an approximation consistent with available

information in the warranty database. Intuitively, its effect might be to bias estimation of the cumulative probability function $F_C(\cdot)$ in the direction of optimism about unit lifetimes. In Section 4, we will see that at least in one example case (where realistically simulated data violate this condition), any distortion in inference caused by creating the log-PL contribution (19) under this condition is small.

To end this section, we note that Majeske, Caris and Herrin (1997) and Lu (1998) also mention the estimation of "sales lag" (the difference between assembly time and delivery time). This aspect is related to our work in that our formulation and estimation of ν_{ij}'s above provides a concrete approach for incorporating sales lag of units into failure time analysis. See Karim and Suzuki (2004), Karim and Suzuki (2005, sec. 10), and Karim (2008) for inference scenarios involving other time lags with warranty data (e.g., lags in the reporting of claims).

3.3 Final log-PL function and inference

Adding the log-PL contributions $\sum_{i=1}^{12} \mathcal{L}_i(\mu, \sigma^2, \alpha, \beta, \theta)$ from the warranty database from Section 3.1 and the log-PL contributions $\mathcal{L}^*(\mu, \sigma^2, \alpha, \beta, \theta)$ from Section 3.2 for assembled units not in the warranty database, we arrive at a final approximate log-PL function $\mathcal{L}(\mu, \sigma^2, \alpha, \beta, \theta)$ for the parameters $(\mu, \sigma^2, \alpha, \beta, \theta)$,

$$\mathcal{L}(\mu, \sigma^2, \alpha, \beta, \theta) = \sum_{i=1}^{12} \mathcal{L}_i(\mu, \sigma^2, \alpha, \beta, \theta) + \mathcal{L}^*(\mu, \sigma^2, \alpha, \beta, \theta). \tag{20}$$

Then for a particular data set we can (numerically):

1. find a vector of parameters $(\hat{\mu}, \hat{\sigma}^2, \hat{\alpha}, \hat{\beta}, \hat{\theta})$ maximizing $\mathcal{L}(\mu, \sigma^2, \alpha, \beta, \theta)$ that can be used as a "maximum PL estimate" of $(\mu, \sigma^2, \alpha, \beta, \theta)$. (This vector provides a "best" fit to the data.)

2. compute the inverse of the 5×5 matrix H of the second partial derivatives of the function $-\mathcal{L}$, evaluated at the PL estimates of the parameters. This provides an estimate of the variance-covariance matrix for the maximum PL estimator. In particular, the square roots of the diagonal elements of H^{-1} correspond to standard errors of

elements of $(\hat{\mu}, \hat{\sigma}^2, \hat{\alpha}, \hat{\beta}, \hat{\theta})$.

3. combine the estimate and the approximate variance-covariance matrix via the delta method to give estimates, standard errors and then confidence limits for interesting functions of $(\mu, \sigma^2, \alpha, \beta, \theta)$ such as the usage rate cumulative density $F_U(u|\alpha, \beta)$ or the cumulative failure time distribution function $F_C(t|\mu, \sigma^2, \alpha, \beta, \theta)$, $t > 0$ for C given in (6). (More details are provided in the Appendix.)

We will see in the next section how the methodology performs in a small simulation study.

4 Simulation Study

For confidentiality reasons, we are not able to present results for actual company data. Hence, both for illustrating our methodology and demonstrating its effectiveness we will use simulated data with characteristics more or less like those of our motivating case. This simulation study is illustrative only, not comprehensive (because an extremely large number of factors impact the generation of the warranty and non-warranty data as will be illustrated in the following). In a real application, however, we suggest using the PL-method to obtain parameter estimates for the data at hand and perform a simulation study, similar to the one next presented, to assess the performance of the method (and the impact of the assumptions made) for the data configuration and problem that the user is addressing.

4.1 Simulation Design

To begin to describe our data simulation, Table 2 provides hypothetical production counts essentially consistent with counts in the motivating case. As before, month 0 means the closing month, month 1 means the month immediately before the closing month, and so on. We simulate data for 32550 units in total.

For each unit, we assign a delivery delay (delivery delay means the time the unit takes to be delivered after being assembled). We will model this with the discrete distribution in Table 3 (that is, again, essentially consistent with the real case).

Simulation using the distribution of Table 3 for 32550 units produced as in Table 1

Table 2: Hypothetical counts of units assembled in each of 30 months.

Month	Count	Month	Count	Month	Count
30	50	19	1500	9	50
29	250	18	1000	8	150
28	1000	17	1000	7	250
27	1500	16	1500	6	500
26	2000	15	1500	5	1000
25	2500	14	2500	4	1500
24	1000	13	500	3	1500
23	500	12	500	2	1000
22	50	11	1000	1	5000
21	50	10	1000	0	2500

Table 3: The probability distribution for delivery delay (months waiting to be delivered after assembly).

Month Delay	Probability	Month Delay	Probability
0	2/30	6	1/30
1	12/30	7	1/30
2	6/30	8	1/30
3	2/30	9	1/30
4	1/30	10	1/30
5	1/30	11	1/30

gives both an assembly month and a delivery month for every unit. Then for each unit we generate values for U and T based on a specific set of parameters. These produce values for A and C using equations (1) and (2) and determine which units produce records in the warranty database and what the repair months are for those units.

Based roughly on the fractions of units of the 16 different cases in the real warranty database as represented in Table 4, we then randomly assign units into the 16 cases. This

leads to simulated data like the real data (i.e., having various configurations of missing information), but for which we know the "truth." We can then see if our methodology can reliably estimate true parameters used to simulate data.

Table 4: Fractions of 16 data types in the real case.

1	0.6	9	0.004
2	0.06	10	0.122
3	0.005	11	0.003
4	0.006	12	0.001
5	0.003	13	0.005
6	0.09	14	0.015
7	0.01	15	0.002
8	0.06	16	0.014

4.2 Estimation Results

To illustrate our methods, we considered one set of parameters and simulated 50 sets of warranty data for this set of parameters; these parameters were chosen based on the estimates from the motivating (actual company) data. Fifty simulation runs were chosen out of computational considerations (recall any evaluation of the log-PL function requires a separate numerical integration for each unit to determine its probability contribution). Software we developed in the `R` statistical system was then applied to the artificial data (for monthly production counts totaling to 32550) and produced maximum PL estimates for the parameters. We summarize results in Table 5.

Average estimates (over 50 simulations) of all parameters are close to the corresponding original parameter values. Also the standard deviations of the estimates are very small. These facts indicate that our method of estimating the parameters is both accurate and precise.

For each simulated data set, we also constructed nominally 95% confidence limits for each parameter, and checked if the confidence limits bracketed the true parameter values.

Table 5: Averages and standard deviations of the 50 parameter estimates for one set of parameters $\{\mu = 5, \sigma = 2, \alpha = 0.45, \beta = 3.4, \theta = -0.5\}$.

Parameter Value	Ave. of Estimates	S.D. of Estimates
$\mu = 5$	5.08	0.04
$\sigma = 2$	1.98	0.01
$\alpha = 0.45$	0.46	0.02
$\beta = 3.4$	3.49	0.03
$\theta = -0.5$	-0.56	0.02

We summarize the numbers (out of 50) of 95% confidence intervals containing the true parameter values in Table 6. Also, we provide the mean and median lengths for the 95% intervals in Table 7. These Tables 6 and 7 provide some evidence that not only does our PL methodology provide effective point estimates for model parameters, but it also provides effective sample/empirical quantification of the quality of those estimates.

Table 6: Numbers (out of 50) of nominally 95% confidence intervals containing the true parameter values.

Parameter	Number of Intervals Covering
μ	46
σ	47
α	49
β	45
θ	48

To demonstrate the importance of incorporating terms in the log-PL for non-failed units (not appearing in the warranty database), we made the maximum PL estimates for the parameters based only on the simulated warranty database. Those are summarized in Table 8. From Table 8, we see that, although the estimates are still stable (have small standard deviations) except for α, the estimates themselves are far from the real parameter values. This supports our assertion that to accurately estimate the distributions of usage rate and

Table 7: Mean and median lengths for the nominally 95% confidence intervals.

Parameter	Mean and Median
μ	0.170
σ	0.048
α	0.102
β	0.110
θ	0.121

failure time for units manufactured and delivered over time, it is necessary to include log-PL contributions for assembled units which do not fail and hence are not represented in the warranty database.

Table 8: Averages and standard deviations of the 50 parameter estimates using only warranty database (excluding additional assembly count information) for one set of parameters. $\{\mu = 5, \sigma = 2, \alpha = 0.45, \beta = 3.4, \theta = -0.5\}$

Parameter Value	Ave. of Estimates	S.D. of Estimates
$\mu = 5$	5.95	0.03
$\sigma = 2$	0.83	0.07
$\alpha = 0.45$	2.58	0.14
$\beta = 3.4$	3.97	0.09
$\theta = -0.5$	-0.82	0.03

Note that the model used in this simulation departs fairly strongly from the approximation $\sum_{j=0}^{i} \nu_{ij} = 1$ used in developing PL terms for units not in the warranty database (as discussed in Section 3.2.2, we formulate PL contributions for non-failed units under the assumption that assembled units will be delivered before the closing time of data collection, but know that this cannot be strictly true of all units). In fact, Table 9 shows that the expected numbers of units produced in months 0 through 12 delivered after the closing time are appreciable. So it would seem that this simulated case is a good test of the extent to which the approximation is likely to degrade estimation and bias estimates of values of $F_C(\cdot)$

Table 9: Expected numbers of units assembled in particular months actually delivered after the closing time.

0	2333	6	83
1	2666	7	33
2	333	8	15
3	400	9	3
4	350	10	33
5	200		

to the low side (making estimates of the failure time distribution consistently optimistic).

To investigate this possibility we make the plot in Figure 1. Pictured there is the actual failure time cumulative distribution function $F_C(\cdot|\mu, \sigma^2, \alpha, \beta, \theta)$ with the 50 estimates $F_C(\cdot|\hat{\mu}, \hat{\sigma}^2, \hat{\alpha}, \hat{\beta}, \hat{\theta})$ and their average. It is clear that any bias in the estimated failure time cumulative distribution function is small. This is at least some evidence that the effect of the approximation $\sum_{j=0}^{i} \nu_{ij} = 1$ can be negligible.

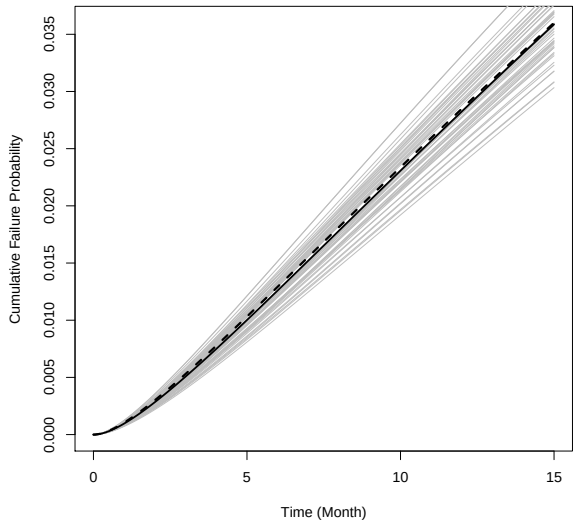

Figure 1: Actual cumulative failure probability function (heavy dashed) compared with the 50 estimates (light solid) and their average (heavy solid).

5 Example Data Analysis

Here we illustrate the practical inference possibilities using our methodology, based on a single simulated data set with total of 32550 products and 2476 warranty database cases.

To see if the non-missing warranty data (i.e., the case 1 units in Table 1 are adequate to describe the data structure, we can plot the usage rate distribution with real parameter values for the purpose of comparing with the histogram of the usage rates of case 1 units in the warranty database. This is illustrated in Figure 2, and the fit is not good. So it is clearly not adequate to estimate the usage rate distribution using only non-missing warranty data.

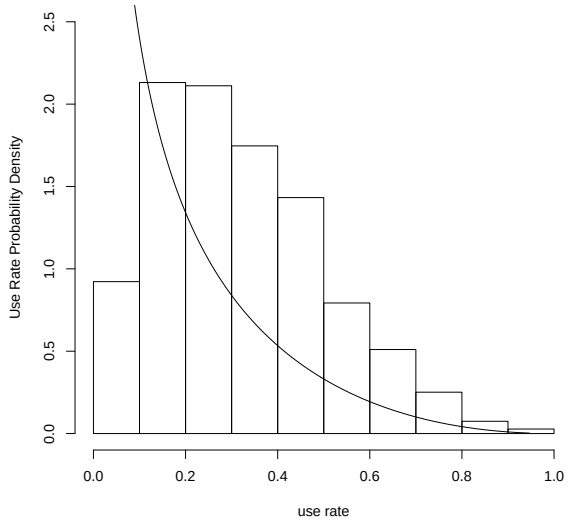

Figure 2: Beta distribution with real parameter values compared to histogram from the simulated case 1 warranty data.

We can also estimate the probability $F_C(t|\mu, \sigma^2, \alpha, \beta, \theta) = P(C \leq t|\mu, \sigma^2, \alpha, \beta, \theta)$ that a unit will fail by at least a time t after being delivered, given in (6). This estimate $F_C(t|\hat{\mu}, \hat{\sigma}^2, \hat{\alpha}, \hat{\beta}, \hat{\theta})$ is plotted as the center curve in Figure 3. Reading from the plot, we can see that our fitted model estimates a 3.3% failure fraction in a one year warranty period. This is consistent with the size of the simulated warranty database in comparison to the total production counts, since a large part of production is near the closing time of data collection, and many units do not yet have 12 months of use at the closing time. In Figure 2, we can also estimate the calendar time, denoted by $F_C^{-1}(p|\mu, \sigma^2, \alpha, \beta, \theta)$, by which time a

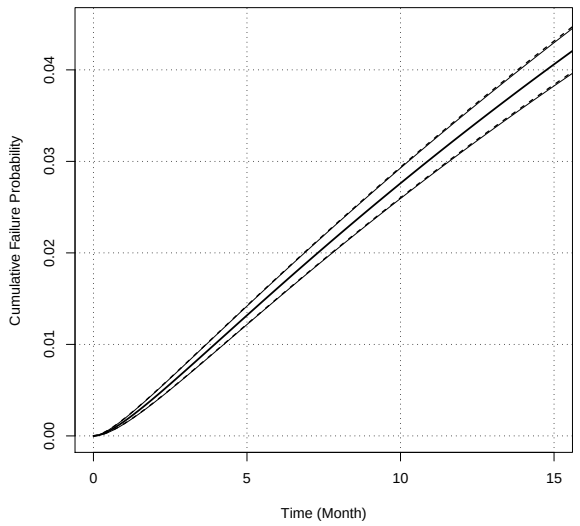

Figure 3: Estimated $F_C(t)$ curve (heavy black in the middle) and sets of 95% confidence limits of failure time for a given cumulative failure probability (solid) and cumulative failure probability for a given time (dashed).

fraction p of units will have failed by selecting a probability value p on the vertical axis and reading off a corresponding time from the estimated probability curve. For example, the time by which $p = 3\%$ of units have failed is estimated to be $F_C^{-1}(0.03|\hat{\mu}, \hat{\sigma}^2, \hat{\alpha}, \hat{\beta}, \hat{\theta}) \approx 11$ months.

Figure 3 also provides some indications of the uncertainty associated with inferences about the cumulative failure probability $F_C(t|\mu, \sigma^2, \alpha, \beta, \theta)$, $t > 0$, (and, reciprocally, the failure percentiles $F_C^{-1}(p|\mu, \sigma^2, \alpha, \beta, \theta)$ for various p) by providing sets of 95% confidence limits. One reads confidence limits for $F_C(t)$ from the solid curves above time t and reads confidence limits for $F_C^{-1}(p)$ from the dashed curves across from p. For example, approximate 95% limits for $F_C(10)$ are $[0.026, 0.029]$ while the approximate 95% limits for $F_C^{-1}(0.02)$ are $[6.9, 7.8]$. These limits are potentially very important in practical inference, and are produced using the method mentioned in Section 3.3 and discussed in more detail in the Appendix. From Figure 3, we see confidence intervals increase in length with t or p, which indicates more uncertainty as time goes by.

6 Conclusion

In this paper we have presented a method to estimate the parameters of a model for warranty data, based on incomplete information. We developed a log-PL function to compute estimates and get confidence limits for parameters and parametric functions, such as the probability of failing within any specific time or the time corresponding to any specific cumulative failure probability.

There are several possibilities for future work and considerations in this area. First (motivated by the real case), we might specify a mixture distribution for the usage rates, for example, as a mixture of two beta distributions. By doing this, we might account for two fundamentally different applications of the units. Second (again motivated by the real case), we might model the possibility that characteristics of a product (as manufactured) change at a known or unknown point in production. Third, we might apply Bayesian methods in place of our fairly ad hoc adjustment of likelihood functions in light of missing information on delivery months, and by doing this, possibly get more effective estimation methods.

Acknowledgements

The authors are grateful to Mike Hamada (Los Alamos National Laboratory) for a careful review of an early draft of this manuscript, which provided us with helpful suggestions and important references.

References

[1] Alam M.M. and Suzuki, K. (2009). Lifetime estimation using only failure information from warranty database. *IEEE Transactions on Reliability*, **58**, 573-582.

[2] Hu, X.J. and Lawless, J.F. (1996a). Estimation of rate and mean functions from truncated recurrent event data. *J. Amer. Statist. Assoc.*, **91**, 300-310.

[3] Hu, X.J. and Lawless, J.F. (1996b). Estimation from truncated lifetime data with supplementary information on covariates and censoring times. *Biometrika*, **83**, 747-761.

[4] Hu, X.J. and Lawless, J.F. (1997). Pseudolikelihood estimation in a class of problems with response-related missing covariates. *Canad. J. Statist.*, **25**, 125-142.

[5] Kalbfleisch, J.D. and Lawless, J.F. (1988a). Estimation of reliability from field performance studies (with discussion). *Technometrics*, **30**, 365-388.

[6] Kalbfleisch, J.D. and Lawless, J.F. (1989). Inference based on retrospective ascertainments: an analysis of the data on transfusion-related AIDS. *J. Amer. Statist. Assoc.*, **84**, 360-372.

[7] Kalbfleisch, J.D., Lawless, J.F., and Robinson, J.A. (1991). Methods for the analysis and prediction of warranty claims. *Technometrics*, **33**, 273-285.

[8] Kalbfleisch, J.D. and Lawless, J.F. (1996). Statistical analysis of warranty claims data, Chapter 9, *Product Warranty Handbook*. Eds. W.R. Blischke and D.N.P. Murthy. New York: Marcel Dekker.

[9] Karim, M.R. and Suzuki, K. (2004). Analysis of field failure warranty data with sales lag. *Pakistan J. Statist.*, **20**(1), 93-102.

[10] Karim, M.R. and Suzuki, K. (2005). Analysis of warranty claim data: a literature review. *International Journal of Quality and Reliability Management*, **22**, 667-686.

[11] Karim, M.R. (2008). Modelling sales lag and reliability of an automobile component from warranty database. *International Journal of Reliability and Safety*, **2**(3), 234-247.

[12] Lawless, J.F., Hu, X.J., and Cao, J. (1995). Methods for the estimation of failure distributions and rates from automobile warranty data. *Lifetime Data Anal.*, **1**, 227-240.

[13] Lawless, J.F. (1998). Statistical analysis of product warranty data. *International Statistical Review*, **66**, 41-60.

[14] Lu, M.W. (1998). Automotive reliability prediction based on early field failure warranty data. *Quality and Reliability Engineering International*, **14**, 103-108.

[15] Majeske, K.D., Caris, T.L. and Herrin, G. (1997). Evaluating product and process design changes with warranty data. *International Journal of Production Economics*, **50**, 79-89.

[16] Meeker, W.Q. and Escobar, L.A. (1998). *Statistical Methods for Reliability Data*. New York: Wiley.

[17] Philips, M.J. and Sweeting, T.J. (1996). Estimation for censored exponential data when the censoring times are subject to error. *J. R. Stat. Soc. Ser. B*, **58**, 775-783.

[18] Philips, M.J. and Sweeting, T.J. (2001). Estimation from censored data with incomplete information. *Lifetime Data Anal.*, **7**, 279-288.

[19] Suzuki, K. (1985a). Estimation method of lifetime based on the record of failures during the warranty period. *J. Amer. Statist. Assoc.*, **80**, 68-72.

[20] Suzuki, K. (1985b). Nonparameteric estimation of lifetime distribution from incomplete field data. *Technometrics*, **27**, 263-271.

[21] Wu, H. and Meeker, W.Q. (2002). Early detection of reliability problems using information from warranty databases. *Technometrics*, **44**, 120-133.

A Details of Inferences Based on the Log-PL

Here we provide more details about inferences for a general real-valued function of the model parameters, $g(\mu, \sigma^2, \alpha, \beta, \theta)$.

Using the maximum pseudo-likelihood estimate (the MPLE)

$$\hat{\boldsymbol{\phi}}^T = \text{argmax } \mathcal{L}(\boldsymbol{\phi}).$$

of $\boldsymbol{\phi}^T = (\mu, \sigma^2, \alpha, \beta, \theta)$ (see Section 3.3), the resulting MPLE for $g(\boldsymbol{\phi})$ is $g(\hat{\boldsymbol{\phi}})$. Let $H(\boldsymbol{\phi})$ be the negative Hessian matrix of $\mathcal{L}(\boldsymbol{\phi})$ from (20). Then $[H(\hat{\boldsymbol{\phi}})]^{-1}$ functions as an estimated

variance-covariance matrix for $\hat{\boldsymbol{\phi}}$. With $g'(\cdot) = (\frac{\partial g(\cdot)}{\partial \mu}, \frac{\partial g(\cdot)}{\partial \sigma}, \frac{\partial g(\cdot)}{\partial \alpha}, \frac{\partial g(\cdot)}{\partial \beta}, \frac{\partial g(\cdot)}{\partial \theta})^T$ and the "delta method" (the propagation of error formula), approximate 95% confidence limits for $g(\boldsymbol{\phi})$ are given by

$$g(\hat{\boldsymbol{\phi}}) \pm 1.96 \times \sqrt{(g'(\hat{\boldsymbol{\phi}}))^T [H(\hat{\boldsymbol{\phi}})]^{-1} g'(\hat{\boldsymbol{\phi}})}.$$

As mentioned in Section 3.3, particular choices of a parametric function $g(\cdot)$ will yield either the probability of failing $P(C \leq t|\boldsymbol{\phi}) = F_C(t|\boldsymbol{\phi})$ within any specific (given) time $t > 0$ (see (6)) or the time $F_C^{-1}(p|\boldsymbol{\phi})$ corresponding to any specific (given) cumulative failure probability $0 < p < 1$. To see this, for a fixed t or fixed $0 < p < 1$, write

$$\tilde{g}(\boldsymbol{\phi}) = F_C(t|\boldsymbol{\phi}), \quad \bar{g}(\boldsymbol{\phi}) = F_C^{-1}(p|\boldsymbol{\phi}).$$

In Section 5, confidence intervals for the quantity $F_C(t|\boldsymbol{\phi})$ were computed using $\tilde{g}$ in the above interval formula, where partial derivatives $\tilde{g}'(\hat{\boldsymbol{\phi}})$ were approximated numerically. Note that, by the implicit function theorem and treating $F_C(t|\boldsymbol{\phi})$ as a function of $(\boldsymbol{\phi}, t)$,

$$\bar{g}'(\hat{\boldsymbol{\phi}}) \;=\; -\frac{\partial F_C(t|\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \left(\frac{dF_C(t|\boldsymbol{\phi})}{dt} \right)^{-1} \Bigg|_{(\boldsymbol{\phi},t)=(\hat{\boldsymbol{\phi}},\hat{c})} \;=\; -\frac{\tilde{g}'(\hat{\boldsymbol{\phi}})}{f_C(\hat{c}|\hat{\boldsymbol{\phi}})},$$

where $\hat{c} = F_C^{-1}(p|\hat{\boldsymbol{\phi}})$ and the density $f_C(\cdot|\boldsymbol{\phi})$ has the form of (4). Confidence intervals for $F_C^{-1}(p|\boldsymbol{\phi})$ can be computed using $\bar{g}$ instead of g in the interval formula.